### Experiment 1

***Method***

##### Participants

##### Fifty-four Dutch native students at Radboud University, Nijmegen, were paid to participate in the experiment. Their average age was 20.9 years (range: 18 - 28) and on average they had taken French classes at school for 3.1 years (range: 1 - 6). Participants were assigned to one of two groups, 27 per group, such that the average amount of years of French classes at school did not significantly differ between the two groups (t[49.77]=0.13). One group was trained with spelling as additional information (+spelling), and the other group was not (-spelling).

##### Materials

A total of ninety-Six words were used as word stimuli in the lexical decision task. Of these words, 24 were bisyllabic with schwa in initial syllable (e.g., la pelouse). The full and reduced variants of these 24 schwa words (e.g., *la pelouse*) served as target stimuli in the lexical decision task. We will refer to these 24 pairs of full and reduced variants as *target word types*. The remaining 72 words served as word fillers in the lexical decision task and they comprised 42 bisyllabic and 30 monosyllabic words. The selection of these 96 words was constrained by the likelihood of participants in the experiment being familiar with these words. It was desirable that target word types be new to participants in order to make sure that any effects arising in lexical decision were due to experimental manipulation rather than to prior experience with these word types. On the other hand, in order to avoid bias during lexical decision, it was necessary that participants knew at least some of the filler words. We therefore based the selection of the 96 words on the outcome of a pretest, in which we assessed the familiarity of 38 native speakers of Dutch, who did not participate in the experiment, with 200 French words. In this pretest, participants were visually presented with each of the 200 words, and they had to indicate whether they knew the word, or not. The 24 target word types in our experiment were those 24 bisyllabic schwa words from the set of 200 words that were least known by the group in the pretest with a high proficiency in French. On the other hand, the 72 filler words in the experiment were the words that were most known by the group in the pretest with a low proficiency. Average lemma frequency of the 24 selected target word types was 3.06 and 8.21 per million according to the *Corpus des livres* and the *Corpus des sous-titres*, respectively in the *Lexique 3* database. Compared to the target word types, the 72 selected filler words were high-frequent with an average lemma frequency of 272.9 and 295.3 per million according to the *Corpus des sous-titres* and the *Corpus des livres*, respectively.

Each of the 24 target word types was associated with its Dutch translation, which was used for the training block. From each of the 24 target word types a matching set of schwa pseudowords was derived for the lexical decision task. Each of these 24 pseudoword sets contained a full bisyllabic variant with schwa in initial syllable (e.g., \**la pessade*) and a reduced variant, in which schwa was absent (e.g., \**la p'ssade*). The full variants of each of these 24 pseudoword sets were created by combining the initial syllable of the full target variants from which they were derived with an existing syllable of French according to the following criteria: the syllable was a possible second syllable of a bisyllabic noun in French with the same gender as the corresponding target word type, the syllable had the same CV-structure as the second syllable of the corresponding target variant (e.g., word: *la pe-louse*, pseudoword: *la \*pe-ssade*), and the syllable did not occur as the second syllable of any other target varian in the experiment. The inclusion of pseudowords that closely matched the target word types ensured that there would not be any systematic difference between target stimuli and pseudoword stimuli and therefore prevented participants from developing trivial strategies during the lexical decision task. The experimental stimuli further comprised 72 pseudowords that had been derived from the filler words, and in which one or two phonemes had been substituted.

A female native speaker of French recorded all stimuli preceded by the definite article *le* or *la* in a sound attenuated booth at a 44.1 kHz sampling rate and 16-bit resolution on a mono channel. For each of the target word types and the corresponding pseudoword types, she produced both reduced and full variants. The reduced variants of the target word types were recorded twice: one recording was used in the training phase, and the other recording in the lexical decision task. Average du­rations of the reduced target variants including the articles were 812ms (range: 556 - 973) and 825ms (range: 562 - 998) in the training and lexical decision task, respectively. Average duration of the full target vari­ants was 858ms (range: 553 - 1026) (*note: Add information about schwa duration?*).

We created four experimental master lists for the lexical decision task, each of which corresponded to a different pseudorandomized sequence of the 192 experimental trials (24 target trials, 72 word filler trials, and 96 pseudoword filler trials). The use of various pseudorandomized sequences was aimed at minimizing effects of trial order on the experimental variables. The four masterlists were pseudoranomized according to the following criteria: at least the first three trials were filler trials, there were no consecutive target trials, and there were at most eight consecutive word or pseudoword trials. For each of the four masterlist, two experimental mirror sublists were created. In each of the two sublists, half of the target stimuli were full variants (e.g. *la pelouse*), and half were reduced variants (e.g., *la p'louse*). Similiarly, half of the 24 pseudowords that had been derived from the target word types were full variants (e.g., \**la pessade*), and half were reduced variants (e.g., *la p'ssade*). Each target word type occurred in its full and reduced variant an equal number of times across lists. However, a target word type that occurred as full variant in one sublist occurred as reduced variant in the other sublist.

##### Design and Procedure

The experiment used a 2 X 2 design with *Spelling* (+spelling vs. –spelling) as a between subject and within word variable, and *Reduction* (reduced vs. full) as a within word and within subject variable. It is important to remember that target word types were learnt by participants only in their reduced variants, and the variable *Reduction* refers to full versus reduced variants in the lexical decision task. Stimulus presentation and data collection of all the tasks were controlled using *Eprime* and *Psychopy* software. All tasks were performed in a sound attenuated booth.

Participants were individually trained and tested on two consecutive days. The training phase (Day 1) consisted of two parts: in the first part, participants had to memorize the translations of the 24 target word types. In each trial, they heard a target word type in its reduced variant and then saw the Dutch translation on the screen. Participants in the +spelling group additionally saw the spelling of the target word type prior to hearing the reduced variant. Each trial started with a visual warning signal that was displayed for 200ms in the middle of the screen. 100ms after the warning signal, a reduced target variant was auditorily presented. 500ms after the offset of the target variant, the corresponding Dutch translation appeared in the middle of the screen for 1500ms. After another 1000ms, the next trial started. During the course of this first part of the training phase, each reduced target variant was presented four times with a lag of at least three intervening trials.

In the second part of the training phase, participants heard the reduced variants of the 24 target word types again. In this part, a trial consisted of the auditory presentation of a reduced target variant, after which participants had to provide the corresponding Dutch translation using a computer keyboard. Again, participants in the +spelling group first saw the spelling of the target word type. A trial started with a visual warning signal that appeared for 200ms on the screen, and which was followed by the auditory presentation of the target variant after 100ms. Participants then had to give the Dutch translation of the target variant using a keyboard that was placed in front of them. There was no time limit for the transcriptions. Participants confirmed their response by pressing the ‘Enter’ button, and they then saw a message on the screen that indicated whether the translation they had provided was correct, or not. If they had not provided a correct translation, the correct translation appeared on the screen. After another press on the ‘Enter’ button, the next trial started. As in the first part of the training phase, each of the 24 target variants was presented four times, and there were at least three intervening trials between repeated trials.

In the test phase (Day 2), participants performed an auditory lexical decision task, in which the word types they had learnt in the training block were repeated either as full or reduced variants. Participants were instructed to decide as quickly and accurately as possible whether the stimulus they heard was a word or a nonword. They were explicitly instructed to press the *nonword* button when they had doubt about the lexical status of a stimulus. Participants indicated their decision by pressing one of two labeled buttons on a response box placed in front of them. The *word* button on the response box was on the side that corresponded to the dominant hand of a participant. Each trial in the lexical decision task consisted of a visual warning signal that appeared for 200ms on the screen. 100ms after the signal, a stimulus was auditorily presented. Participants then had to make their decision. 200ms after participants had pressed a button on the response box, the next trial started.

##### Additional assessments

In order to ensure that any differences between both participant groups found in the lexical decision task were due to experimental manipulation rather than due to differences in proficiency in French, it was necessary that both participant groups did not significantly differ in their respective proficiency in French. We therefore assessed proficiencies of all participants by means of the LexTale test of proficiency for French (Brysbaert, 2013) and by means of a questionnaire in which participants had to provide a self-estimated score of proficiency in reading, writing, speaking and listening to French on a scale from 1 to 6.

The LexTale test of proficiency for French is a visual lexical decision task including 56 French words and 28 French looking nonwords. We calculated the score for a participant by subtracting two times the number of nonwords that she incorrectly identified as words from the number of words that she correctly identified as words (Nwords selected - 2 \* Nnonwords selected) . The same formula is used by Brysbaert (2013) for calculating the scores of 289 students with French as L2. A comparison of the scores of the participants in our experiment with the students tested in Brysbaert (2013) allowed us to get an impression of the overall proficiency of the participants in our experiment. Their average score based on the mentioned formula was -2.03 (range: -14 to 12), which corresponds to the first decile of scores from the 289 students with French as L2 tested in Brysbaert (2013). This suggests that the participants in our experiment were relatively low-proficient (*note:* *I maybe have to elaborate on this a bit.*) Importantly, the scores of participants in the +spelling group did not significantly differ from the scores of participants in the –spelling group (+spelling: -1.30, –spelling: -2.78; *t*[49.73]=0.87).

We took the average score in reading, writing, speaking and listening to French of a participant provided in a questionnaire as an indicator of the level of proficiency of that participant. The average proficiency of all 54 participants based on this score was 2.05 (range: 1.0 - 3.75). Importantly, there was no significant difference between the scores of participants in the +spelling group and the scores of participants in the –spelling group (+spelling: 2.12,   
–spelling: 1.98; *t*[50.60]=0.72).

The assessment of participants’ proficiency in French described in the previous two paragraphs served to ensure that proficiency was not a confounding factor in our analyses of the effect of spelling. A further potential for confound could be a difference in learning success between participants in the +spelling and participants in the –spelling group. In order to assess whether participants of both groups learnt the words presented to them in the training equally well, we conducted a vocabulary task prior to the lexical decision task on Day 2. In this task, each of the 24 target variants that participants had been exposed to in the training phase on Day 1 were again presented. In each trial, participants had to identify the correct translation of a given target variant out of four possible translations that were displayed on the screen. A trial started with a visual warning signal that was displayed for 200ms on the screen. 100ms thereafter, participants heard a target variant and saw four Dutch words on the screen, with one word being the correct translation and three words being translations of other target word types. The four words were horizontally arranged with the position of the correct translation randomly determined by computer controlled randomization. During the course of the vocabulary test, all 24 translations of the 24 target word types appeared as possible choices four times. Participants indicated their decision by mouse clicking on the translation of their choice. 200ms after the click, the next trial started.

***Results***

*Vocabulary test*

(*Note: Numbers in this paragraph are not up to date.*) The overall hit rate in the vocabulary test was 87.4%. Participants in the +spelling group identified the correct Dutch translations in 87.3% of the cases, and in the –spelling group in 87.5% of the cases. According to a chi-squared test for independence, there was no correlation of hit rates and *Spelling* group(χ2 = 0; df = 1; ptwo-tailed= 1).

*Lexical decision*

(…)

### Experiment 2

***Method***

##### Participants

##### Fifty-four Dutch native students at Radboud University, Nijmegen, were paid to participate in the experiment. None of them were participants in the preceding experiment. Their average age was 20.8 years (range: 18 - 29) and on average they had taken French classes at school for 3.0 years (range: 2 to 6). Participants were assigned to one of two groups, 27 per group. The average amount of years of French classes at school did not significantly differ between the two groups (t[48.10]=0.07). One group was trained with spelling as additional information (+spelling), and the other group was not (-spelling).

*Materials*

The same experimental lists as in Experiment 1 were used. In addition to the stimuli used in Experiment 1, a second recording of each of the full target variants was used for the training phase. The other recordings were those that were also used in Experiment 1, and they were used as stimuli in the lexical decision experiment.

*Design and Procedure*

The design and procedure were identical to Experiment 1 with the exception that all participants exclusively heard the full variants of the 24 target word types in the training phase.

*Additional assesments*

As in Experiment 1, we assessed participants’ proficiency in French. The average score of participants in the LexTale proficiency test for was of -3.76 (range: -15 to 6), which corresponds to the first decile of scores from 289 students with French as L2 tested in Brysbaert (2013). The scores did not significantly differ between the +spelling and the –spelling group (+spelling: -2.93, –spelling:   
-4.59; *t*[50.02] = 1.12). The average score based on the questionnaires was 2.06 (range: 1.0 to 3.75), and it didn’t significantly differ between groups, either (+spelling: 2.07, -spelling: -2.04; *t*[51.60] = 0.18).

The learning success of participants was assessed by the same vocabulary task that was used in Experiment 1.